Review

A Survey on Machine Learning Approaches for Personalized Coaching with Human Digital Twins

Harald H. Rietdijk [1,*], Patricia Conde-Cespedes [2], Talko B. Dijkhuis [1], Hilbrand K. E. Oldenhuis [1] and Maria Trocan [2]

[1] Lectoraat Digital Transformation, Hanze University of Applied Sciences, 9747AS Groningen, The Netherlands; t.b.dijkhuis@pl.hanze.nl (T.B.D.); h.k.e.oldenhuis@pl.hanze.nl (H.K.E.O.)

[2] Lisite, Institut Supérieur D'électronique De Paris (Isep), 92130 Issy-les-Moulineaux, France; patricia.conde-cespedes@isep.fr (P.C.-C.); maria.trocan@isep.fr (M.T.)

[*] Correspondence: h.h.rietdijk@pl.hanze.nl; Tel.: +31-6-51-91-6262

Abstract: Human Digital Twins are an emerging type of Digital Twin used in healthcare to provide personalized support. Following this trend, we intend to elevate our virtual fitness coach, a coaching platform using wearable data on physical activity, to the level of a personalized Human Digital Twin. Preliminary investigations revealed a significant difference in performance, as measured by prediction accuracy and F1-score, between the optimal choice of machine learning algorithms for generalized and personalized processing of the available data. Based on these findings, this survey aims to establish the state of the art in the selection and application of machine learning algorithms in Human Digital Twin applications in healthcare. The survey reveals that, unlike general machine learning applications, there is a limited body of literature on optimization and the application of meta-learning in personalized Human Digital Twin solutions. As a conclusion, we provide direction for further research, formulated in the following research question: how can the optimization of human data feature engineering and personalized model selection be achieved in Human Digital Twins and can techniques such as meta-learning be of use in this context?

Keywords: personalization; human digital twin; machine learning; healthcare; artificial intelligence; coaching

1. Introduction

The concept of Digital Twins (DTs) originated in the 1970s when NASA introduced mirrored systems in its Apollo Program to monitor spacecraft during missions. It was in their 2010 and 2012 Technology Roadmaps that they first coined the term Digital Twin [1]. Since then, the concept of using a simulated environment to model a real-world system has proven to be a highly useful and powerful tool. Following the latest developments in industry and technology, the concept has evolved from systems using exact physical replicas to virtual systems with digital copies of the physical world, Digital Twins. Because of this constant evolution, there is no unified definition of Digital Twins [2]. However, the consensus is that Digital Twins consist of the following three minimum components. (i) The physical object or physical twin, the object that is monitored. (ii) The digital duplicate, or Digital Twin, where a model of the monitored object is generated and interventions or optimizations can be evaluated. (iii) The bidirectional data flow between these two [3], which ensures that the physical and digital twins remain in sync.

Today, Digital Twins are applied in various industries, including manufacturing, human–cyber systems [4], healthcare, aerospace, smart cities, and business [5]. They can be used, among other things, for data visualization, to optimize the performance of physical objects or systems in real time, to predict the future behavior of their physical twins, or even in online sales [6].

Since it is essential for a Digital Twin to emulate the state of a physical twin, it is important that the Digital Twin is fed with all the relevant information from the physical twin and its surroundings. In systems where human data is processed, as is common in healthcare, data collection can be complicated. First, it can be challenging to collect a robust number of samples. This can be due to ethical and privacy considerations [7] that must be taken into account when handling personal data or the fact that the research focuses on a specific group of patients or a select group of participants. The result of these challenges is that research and experimentation often have to be performed on datasets with relatively small sample sizes, making it more difficult to reach general conclusions, prevent overfitting of generated models, and achieve sufficient statistical accuracy [8]. In some cases, historical datasets can be used to complement acquired data [9]. If this is not a viable solution to overcome this problem, simulated data can be used [10].

Another challenge lies in the fact that humans are complex organisms and have proven to be difficult to integrate into digital systems. Not only can it be technically complicated to acquire the desired readings, but consistent and complete data acquisition also depends on the participant's or patient's willingness to be monitored. Disloyalty to the research program can lead to data gaps, which require a systematic approach to data imputation [11].

For these reasons, in the healthcare sector, Digital Twins have traditionally been used primarily for predictive maintenance of medical equipment to improve its performance [12,13]. Additionally, Digital Twins have been utilized in hospital management to enhance patient care coordination. A special kind of Digital Twin is the Human Digital Twin (HDT). Originally, these refer to generic virtual models that represent digital replicas of organs or emulations of the human body as a whole. The models used in these HDTs are based on data collected from multiple patients, resulting in digital duplicates that simulate expected generic behavior. Because of this, these applications cannot be classified strictly as personal HDTs.

With the introduction of devices that interact and communicate over the Internet, the Internet of Things (IoT), Big Data, Cloud Computing, and developments in Artificial Intelligence (AI), and specifically Machine Learning (ML), the availability of data and data processing has improved [14]. Specifically, innovations in wearables have made a wider range of measurements available, providing personalized information on individual activity, physical state, and even emotional state. Together with the improved quality of the collected data, this has opened up all sorts of new opportunities for the application of personalized Human Digital Twins in healthcare where predictions and interventions are based on the results of models generated using only personal data.

With these personalized HDTs or patient Digital Twins, the focus lies more on personalizing and individualizing treatments and therapy aimed at changing the behavior of patients or participants [15]. Personalized models enable the achievement of precision healthcare, facilitating proactive patient care, early disease diagnosis and detection [16], and the possibility of preventive care.

Taking these developments into account, our research group decided to investigate the possibilities of elevating our personalized physical activity coaching application, the Virtual Fitness Coach (VFC) platform [17,18], to the level of a personalized HDT. In the first step to achieve this, we focused on evaluating the performance of machine learning techniques used to perform the predictive tasks, we established that the efficiency and usability of

personalized applications such as the VFC platform depend not only on the quality and completeness of the data being used but also on the way this data is processed. This motivated us to determine the state of the art of the application and selection of ML techniques used in the AI cores of HDTs that perform the predictive and analytical tasks.

The rest of this paper is organized as follows: In Section 2, a further motivation for this survey is presented. Next, in Section 3, we discuss the context, definitions, and direction of the survey. Related surveys are given in Section 4. In Section 5, the main findings of existing research on the topic are presented. Finally, the conclusion and suggestions for further research are discussed in Section 6.

2. Motivation

2.1. The Virtual Fitness Coach, an Example of a Personalized Application

The purpose of the VFC platform is to provide coaching to individuals who work sedentary jobs, promoting a healthy lifestyle and physical activity during the workday. Based on step data collected using wearable devices, the platform predicts whether a daily goal will be achieved and then sends personalized coaching if necessary. The evaluation of step data can be performed using personalized models that have been trained with individual data or generalized models trained with data from all participants.

As with other examples in this domain, the extent to which the application goal is achieved, that is, the quality of the coaching, strongly depends on the way the data is processed and the accuracy of the predictions made, i.e., the quality of the coaching stands or falls with the performance of the machine learning implementations. Inappropriate or poorly timed interventions will cause the participant to lose interest in using the platform [19,20].

2.2. Preliminary Results

Our research in [21] focused on personalizing ML implementations on the VFC platform. Using general guidelines for selecting ML classification algorithms based on the data structure and problem characteristics, eight different algorithms were tested, AdaBoost (ADA), Decision Trees (DTC), k-Neighbors (KNN), Logistic Regression (LR), Neural Networking (NN), Stochastic Gradient Descent (SGD), Random Forest (RF), and Support Vector Classifiers (SVCs). This selection includes algorithms that have the potential to perform well on the given problem but also provides different algorithms for classification. The results show that the performance evaluation of the selected algorithms, based on the complete set of participants, does not give the same results as the performance evaluation based on a single participant. First, the performance metrics for personalized models are significantly better than those for general models, and secondly, for personalized models, the highest accuracy is obtained by the DT and RF models, whereas for general models, the ADA, NN, SGD, and SVC models show better results.

To illustrate these findings, some of the results presented in this article are reproduced in Figure 1. In this graph, the accuracy and F1-scores in the test sets are compared for the models generated using eight different classifiers. The first bars represent the accuracy and F1-score results for the models generated using data from all participants. The second bars are the results of the personalized models for each participant. In this case, for each participant, eight different models were fitted using a cross-validated grid search with the selected classifiers, resulting in eight models per participant. In all cases, a 30–70% split was used for the test and training sets. The precision and F1-score reflect the model's capacity to predict the probability that a participant will achieve the step goal at the end of the day based on the hour of the day and the activity recorded up to that hour for the participant.
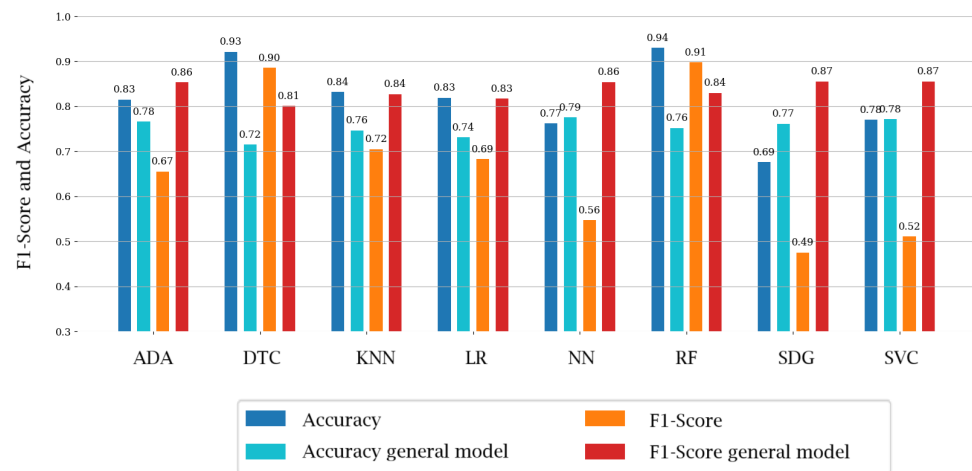
Figure 1. Algorithm F1-score and accuracy comparison using the average of the results obtained by using personalized models and general models, as presented in [21].

Due to the stochastic nature of some algorithms used, the results presented here differ slightly from those in the original paper. Furthermore, a more recent version of Scikit-Learn had to be used, which resulted in improved results for the k-Neighbors, Neural Networking, and Stochastic Gradient Descent classifiers. These differences do not alter the interpretation of the results, namely that not only does personalized modeling yield higher performance scores, but we also observe that the ranking of classifiers is significantly different. The results discussed in [21] also show that the optimal choice of the ML algorithm and the settings of the corresponding parameters differ from individual to individual, underlining the fact that a personalized approach should be considered when using data from several individuals, obtained with the use of wearables.

2.3. Research Goals

In our research presented in [21], model fitting was performed using all algorithms on both the complete dataset and individual datasets. This is a very time-consuming process that should be optimized. Reflecting on these findings and the intention to achieve an HDT implementation of the VFC platform, several questions were raised. What do we know about the performance and selection of ML algorithms in similar applications? What are the arising challenges in the context of personalized HDT applications where the data used is generated by human actions or other human sensor data? Are there more efficient ways to achieve model personalization, for example, based on the characteristics of the individual dataset? To the best of our knowledge, there are currently few answers to these questions. Therefore, to obtain an overview of what is known and to guide our future research, we present a survey on the use of Machine Learning approaches for personalized coaching with Human Digital Twins.

3. Context, Research Questions, and Search Strategy

3.1. Context

As with DT definitions, there is no uniform definition of a HDT. We use the following definition. A Human Digital Twin is any application that uses personal and environmental data to model the behavior of individuals, enabling personalized coaching or interventions.

Although there are many versions of the HDT architecture [5,14,22,23], in general, the following components or layers can be identified [24]. A data acquisition layer, where IoT sensor data, wearable device readouts, stored data, or manual input are collected. A storage layer that usually includes data denoising and preprocessing modules. An analysis and computation layer. In this layer, ML is used to evaluate input and create predictive

models. And finally, there is a user-interaction layer. In this layer, the state simulation is implemented, and communication with the physical twin or other stakeholders is realized.

As an example, the proposed architecture of the HDT version of our VFC platform [17] is shown in Figure 2. This platform communicates with the application that participants install on their mobile phone and can be monitored using an administration dashboard. Neither of these components is presented in this figure. The data acquisition layer has three components. During the registration process, participants submit a selection of personal data using the mobile application. During the participation period, data from a wearable device is collected using the API provided by the wearable device manufacturer. In addition, weather and location data will be collected using the mobile phone and available weather services.
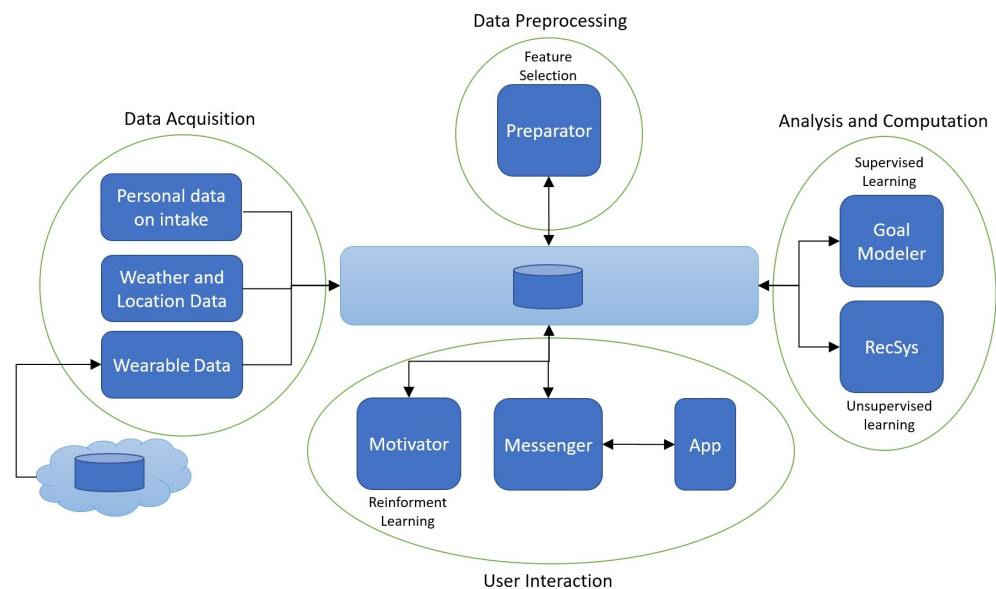

Figure 2. Virtual fitness coach components.

In the preparator, the wearable data is transformed into a format that can be used by the goal modeler. In this layer, it is also possible to apply feature selection algorithms to optimize the input of the motivator. In the analysis and computation layer, three processes take place. (i) Using supervised learning and data collected during the intake period, classification models are generated to predict the likelihood of participants achieving their daily activity goal. (ii) Daily activities are monitored and fed into the models to obtain predictions. (iii) The predictions are evaluated to detect model drift and participant adaptation, in which case models will be regenerated or reassigned.

If necessary, the recommendation system can trigger an intervention that is generated in the user-interaction layer. Here, the motivator selects an intervention based on the personal data collected on intake and current state. This will be communicated through the messenger service to the mobile app. After that, the participant can provide feedback that will be used as input for the reinforcement learning component of the motivator.

3.2. Research Questions

This example clearly shows that ML can be applied in several components of HDTs. In this study, we will focus on establishing the state of the art in the selection and optimization of ML algorithms in personalized HDT applications in healthcare. To establish this state of the art, the first component to consider is machine learning algorithms, their applications, and current guidelines in algorithm selection. The second component is the application of ML in HDT applications within healthcare. In this survey, we are interested in evaluating the performance of personalized ML implementations of these HDTs. The last

component is the human behavior datasets, their characteristics, and their relevance in the process of selecting the ML algorithm. From these three components, we can formulate the following questions that need to be answered to arrive at the state of the art.

1. What is the general approach used when selecting and evaluating ML algorithms?
2. Which ML algorithms are relevant in personalized HDTs?
3. How is the performance of ML implementations evaluated in HDTs?
4. What do we know about the influence of the characteristics of human (behavior) datasets on the performance of different ML algorithms?

3.3. Search Strategy

To answer the questions formulated above, the following steps were performed. A graphical representation of this process is given in Figure 3.

1. Defining search terms and inclusion criteria by evaluating the research questions.
2. Collecting articles by selecting the right sources and performing search queries.
3. Screening and selection of relevant articles by applying the inclusion criteria.
4. Extending the search through the inclusion of relevant references. Repeat step 3.
5. Analyzing the findings to answer the research questions and presenting the results, this is conducted in Section 5.
6. Discussing and interpreting the results, this is conducted in Section 6.
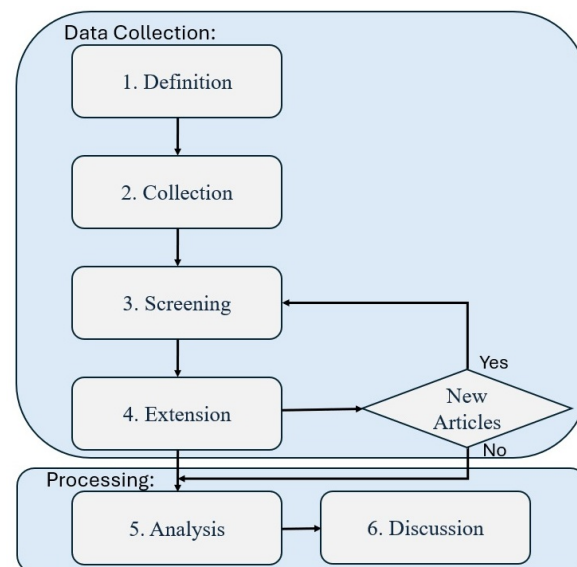
Figure 3. Research process.

3.3.1. Definition

In Table 1, the search terms that were used for the four questions are given.

Table 1. Search terms used per question.

Question	Search Terms
1	machine learning algorithms *followed by* overview, selecting, optimizing, performance evaluation metrics.
2	Digital Twin healthcare, Human Digital Twin *followed by* algorithms, personalized healthcare, coaching, machine learning.
3	Human Digital Twin *followed by* performance evaluation, optimization.
4	human Digital Twin meta learning, machine learning *followed by* human behavior data sets, data set characteristics evaluation, data set meta learning.

Studies of any design were included in this survey. For the first two questions, we excluded articles that did not include a comparison between different algorithms when evaluating their results. For questions 3 and 4, only papers that gave technical details on the implementation of the computational component were included.

3.3.2. Collection, Screening, and Extension

We utilized the search engines of Hanzemediatheek, 136 underlying databases, and those of Sorbonne University, comprising 236 databases. Further use was made of Google Scholar, with an undisclosed number of databases. Searches were conducted between September 2023 and September 2024, and no language or date restrictions were applied. All results found were retrieved in the BibTeX format. This data was transformed into JSON to import it into Excel, where screening and eligibility checks were performed. In this process, all articles that did not appear in a publication related to computer science, were not published in English, or did not meet the requirements mentioned in the previous section were excluded. The remaining articles were accessed online to be evaluated for eligibility. Articles found eligible were imported into the Mendeley Reference Manager, where further processing was performed. The results of this process can be seen in Figure 4.
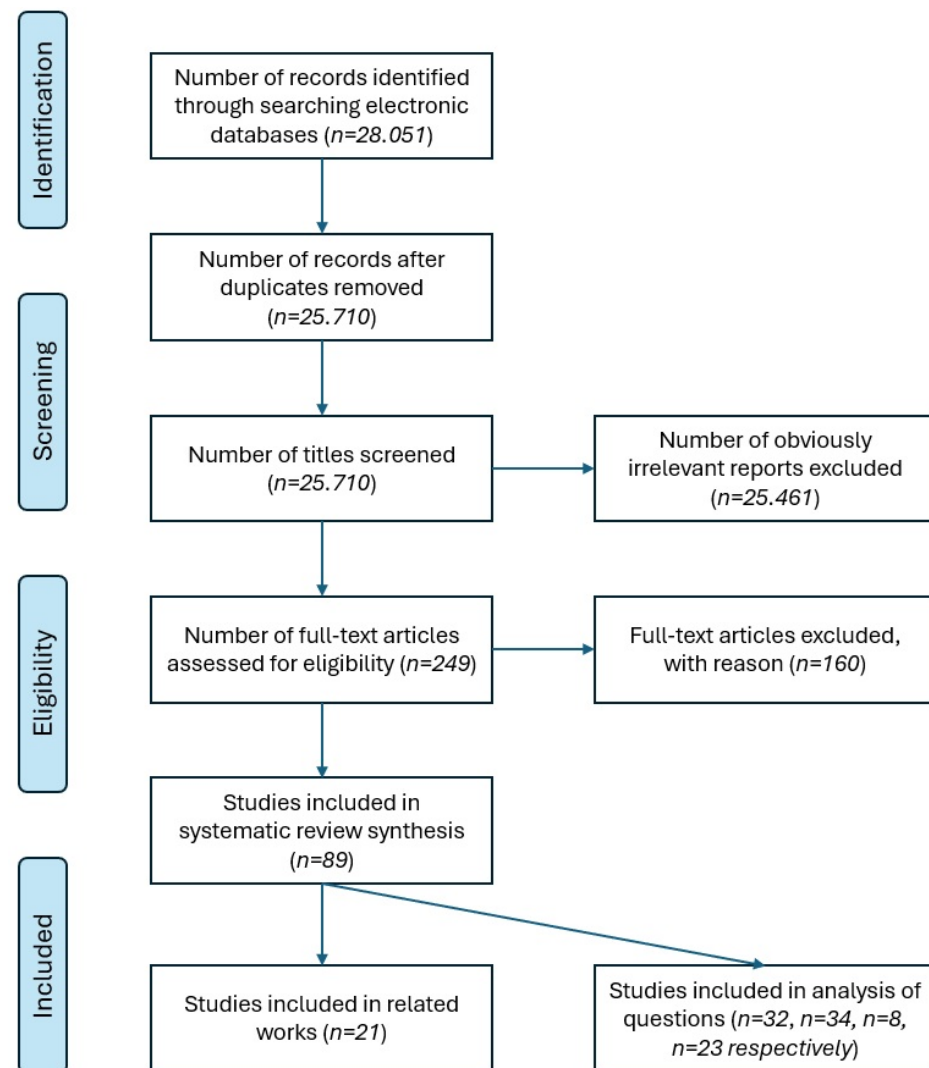

Figure 4. Search results.

4. Related Works

4.1. Surveys on Machine Learning Applications

Alzubi et al. [25] provide a comprehensive introduction to ML processes and algorithms. A framework for categorizing ML algorithms into different paradigms is presented, and a selection of the most common algorithms is categorized accordingly. One of the challenges identified in this paper is the need for large volumes of data to improve accuracy and efficiency, as well as the sparsity of data in fields such as healthcare.

A much more elaborate exposition can be found in Sarker [26]. The aim of his article is to identify the challenges and research directions in the application of machine learning by looking at the components that define the field, namely, data, algorithms, and real-world applications. Based on the type of Machine Learning and the kind of task that must be performed, the algorithms are divided (see also Section 5). After discussing ten popular application areas, their first conclusion is that *the effectiveness and the efficiency of a machine learning-based solution depend on the nature and characteristics of the data and the performance of the learning algorithms.* Another conclusion they draw is that *selecting a proper learning algorithm that is suitable for the target application is challenging. The reason is that the outcome of different learning algorithms may vary depending on the characteristics of the data. Selecting the wrong learning algorithm would result in producing unexpected outcomes that may lead to loss of effort, as well as the model's effectiveness and accuracy.*

In [27], Mahesh provides an introductory overview of the most widely used algorithms. Although general guidelines are given on when different algorithms are applicable, no systematic approach is given. For example, about the transductive support vector machines (TSVMs), they say: *Around it, there has been mystery because of lack of understanding its foundation in generalization.*

These three articles provide general guidelines for the selection and classification of ML algorithms. In [27], only a limited selection of algorithms belonging to the supervised and unsupervised learning paradigms are discussed, whereas in [25], the applicability and limitations of algorithms belonging to all paradigms are presented. In [26], the different algorithms are discussed according to the task they perform. Together, these surveys provide a starting point for answering our first research question.

A good example of a survey with a more specific focus on a healthcare problem is the work by den Hengst et al. [28]. In their work, they discuss the applications of reinforcement learning (RL) to personalize digital systems. Among other things, they provide an overview of the evaluation strategies employed in the solutions used in the studies examined. In discussing their results, they state the following: *Similarly, the results show no increase in the relative number of studies with a comparison of approaches over time. These may be signs that the maturity of the field fails to keep pace with its growth. This is worrisome, since the advantages of RL over other approaches or between RL algorithms cannot be understood properly without such comparisons. Such comparisons benefit from standardized tasks. Developing standardized personalization datasets and simulation environments is an excellent opportunity for future research.*

Early work on the use of machine learning algorithms in healthcare can be found in Yoo et al. [29]. In their work, they provide guidelines on how to utilize various algorithms in the biomedical and healthcare fields, depending on the specific task that needs to be performed. After identifying a number of problems and challenges with data mining in healthcare, their final conclusion is that *An ideal data mining package should (1) support intelligent data preprocessing that automatically selects and eliminates data for the purpose of data mining and uses domain knowledge for various data processes, and (2) fully automate the knowledge discovery process so that it understands and utilizes existing knowledge in data mining processes for better knowledge discovery.*

In both articles, a number of ML algorithms are discussed. Ref. [28] focuses specifically on algorithms used in RL, while [29] lists a number of classification and clustering algorithms. The findings of these surveys show general guidelines for the selection and applicability of the algorithms discussed, which gives us part of the answer to the second research question we formulated.

4.2. Surveys on Digital Twins

As can be seen in the survey conducted by Semeraro et al. [30], there is abundant material on DTs, as well as surveys on the subject. More general surveys on technologies and applications can be found in the works of, for example, Barricelli et al. [3] or Fuller et al. [2]. Surveys with a more general discussion of (H)DTs can be found in [31–36]. In these reviews, the focus lies more on the history of HDTs, the architecture and components that make up an HDT, and the ethical questions surrounding the use of HDTs.

The work of Gámez et al. on DTs in coaching [37], and the study by Minerva et al. on the IOT context of DTs [38], are also worth mentioning. In the first report, an overview of the algorithms, sensors, and platforms most used in DT applications for coaching can be found. In the second one, all the aspects that should be considered in setting up a DT architecture are discussed. Although the focus lies on industry, one of the scenarios examined in depth is that of a digital patient.

In our study, we focus on the healthcare applications of HDTs and the implementation of machine learning in this context. Several surveys related to this area have been found. In the work of Mihai et al. [5], a section is dedicated to the different machine learning algorithms that are used in DT implementations. They find that the use cases and services implemented generally determine the choice of model, but they also state, *However, ML models, and in particular DL (Deep Learning) methods, are generally perceived as black boxes* .

When looking at surveys on Human Digital Twins, a general overview can be found in the article by Lin et al. [12]. Five different types of human data are defined, and HDTs are grouped based on two criteria. The first is the type of modeling, which includes human body/organ modeling and human behavior modeling, with subdivisions such as activity, social interaction, and lifestyle. The second is the application domain, healthcare, industry, or daily life. Two surveys that look at the architecture of HDTs in healthcare are [14,23]. Both articles end with a list of open challenges. In both lists, the need for a better understanding of ML frameworks is mentioned. Furthermore, [23] underscores the challenges that arise due to the complexity of data collection, preparation, and processing, and [14] states that *The deployment of twins for healthcare is mainly affected by two factors, available computing power and latency.*

Although these surveys provide a clear overview of the architectures used in HDTs, they do not investigate the selection and application of ML algorithms in detail, which is required to answer the last three research questions we formulate.

5. Principal Results

In this Section, the answers to the four research questions will be presented. Each subsection focuses on one of the questions and will be preceded by a categorized listing of the relevant articles.

5.1. Findings on Machine Learning

The aim of this section is to answer the first research question, What is the general approach used when selecting and evaluating ML algorithms? As can be seen in Table 2, the selection process found a total of 47 articles that had related studies or presented a relevant study on ML. Of the 32 articles that were found on ML, 9 discussed selection

methods for ML algorithms and 11 focused on ML performance. A further five articles touched on both subjects.

Table 2. Articles found on machine learning.

Category		Publications
Machine learning algorithms	general	[39–45]
	selection	[2,25–27,46–50]
	performance	[8,29,51–59]
	both	[28,60–63]

The articles listed in this section provide general information and definitions of selection criteria. Most often, algorithms are selected since they are known to perform well under certain general conditions. More specific methods for selecting ML algorithms are found in the publications selected to answer the final question. These will be discussed in Section 5.4.

Since there are some minor differences in the way ML algorithms are categorized, this section presents the definitions and concepts that will be used throughout the rest of this study. Since these are general definitions, readers with advanced knowledge of the subject may want to skip to the next Section.

The basic concept of AI is to use a set of rules and available information to make a decision (reasoning). For example, in game playing, the result of possible next moves can be scored using a given heuristic, resulting in the best move based on the highest score. The concept of AI can be enhanced by integrating learning with reasoning. The term "Machine Learning" was introduced in 1959 by Arthur Samuel [39] in precisely this context, when he discussed machine learning procedures using the game of checkers. Learning is achieved when decisions are not made based on previously coded rules, but rather based on experience. A widely used definition of machine learning, expressing this idea, was given by Mitchell [40] in 1997.

> A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

5.1.1. Algorithm Selection and Classification

The way learning is achieved, i.e., the algorithm used, can be categorized into learning model paradigms. Using the broadest definition, four paradigms, as shown in Table 3, are generally identified: Supervised Learning (SL), Unsupervised Learning (USL), Semi-Supervised Learning (SSL), and Reinforcement Learning [26]. As can be seen from the definitions, this classification of learning models is based on the labeling and rules known from the available information, as well as the way feedback is given or processed. Our findings show that all paradigms are used in the context of HDTs.

In some cases, multiple algorithms are used, which do not have to be of the same paradigm. In these cases, using more specific definitions and criteria, looking, for example, at the character of the feedback, or the structure and architecture of the algorithm, other paradigms can be defined, such as multitask learning, evolutionary learning, ensemble learning, instance-based learning, dimensionality reduction algorithms, hybrid learning, neural networks, and deep learning [2,25,27].

The classification of ML algorithms can be refined by looking at the objective of the algorithm [25,26]. These objectives can be divided into three categories: Data, Solution, and Mixed. The two types of problems that belong to the data category focus on improving

the processing of the data by obtaining information about the structure of the input data or modifying the data. These types include dimensionality reduction and feature learning, as well as association rule learning.

Table 3. Machine learning paradigms.

Paradigm	Description
Supervised Learning	In Supervised Learning, a set of labeled training data is used to learn a mapping between the input and output variables. The predictions made by the algorithm are based on historical data.
Unsupervised Learning	In Unsupervised Learning, the aim is to discover hidden structure in unlabeled data with minimal human supervision. It is used to identify existing patterns in the data that have not been previously identified and to create rules based on them.
Semi-Supervised Learning	In Semi-Supervised Learning, both labeled and unlabeled data are utilized to learn from the structure present in the unlabeled data and use it to improve the model's accuracy. It is used when a large amount of unlabeled data is available and only a small amount of labeled data is available.
Reinforcement Learning	In Reinforcement Learning, rewards and punishments are used to teach an AI agent how to behave in an environment. Learning results from interactions with the environment, rather than from labeled datasets. By taking actions that lead to the highest reward, while avoiding actions that lead to punishment, the AI agent maximizes the reward it receives from its environment.

The five types that focus on processing the data into new information form the solution category: Classification analysis, which assigns a class label to an input data point based on its features. Regression analysis models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Cluster analysis involves exploring and analyzing the structure of a dataset, identifying patterns, and discovering natural groupings of data points. Reinforcement learning uses rewards and punishments to teach an AI agent how to behave in an environment. Anomaly detection involves identifying data points that are unusual or out of the ordinary compared to the rest of the data.

The last category mentioned is the mixed category, which contains problems involving artificial neural networks and deep learning, where a combination of the other categories is used. This classification is summarized in Table 4.

Table 4. Classification of ML techniques based on the objective.

Objective	Type
Data	Dimensionality Reduction and Feature Learning
	Association Rule Learning
Solution	Classification
	Regression
	Clustering
	Reinforcement Learning
	Anomaly Detection
Mixed	Deep Learning

5.1.2. Evaluation and Improvement of Performance

Once implemented, the evaluation of algorithm performance can be divided into operational performance, measured in speed and training time, and analytical performance, whose metrics depend on the type of prediction. When the prediction can be classified as true or false, accuracy, precision, recall, specificity, F1-score, and Area Under the Curve are used. These metrics are calculated using the number of true and false positives and negatives [51,56]. When considering regression problems, metrics such as the fitting error, model accuracy error, root mean square error, variance interpretation rate, coefficient of determination, and Pearson correlation coefficient can be used.

Although operational performance can be important, most evaluation efforts focus on analytical performance. There are several approaches to improving the performance of algorithms. The most common is to use training data or data acquired during an evaluation period to compare the performance of several algorithms. Generally, this requires a significant amount of processing time, research time, and effort. Therefore, these evaluations are conducted at a single point in time, typically at the application's conception. This approach is particularly useful when examining systems where behavior remains constant over time, meaning the correlation between data features and observed results remains unchanged. This approach can be used, for example, in developing models to predict the prevalence of influenza-like diseases [52], or in industrial situations when developing smart control applications.

Another approach to improving analytical performance is to divide the problem to be solved horizontally or vertically. With horizontal division, we refer to the implementation of several layers of algorithms or parts thereof. Vertical division, similarly, refers to the parallel implementation of components. Deep learning, Ensemble learning, and Multi-Task learning are examples of this approach. The idea behind this approach is to simulate the human thinking process, which operates at multiple levels, rather than a simple one-layer algorithm. The aim is to be able to understand more complex problems, such as facial recognition or automated text generation [54]. This approach is also used in the analysis of time series for anomaly detection or forecasting of regression results [52,55].

Hybrid learning is a similar approach to improving the performance of an ML solution. The main advantage of hybrid learning is that it combines the strengths of SL and USL. This allows for more accurate predictions and better generalization of the model; An example focused on mortality prediction in COVID-19 cases can be found in [59]. Additionally, hybrid learning can be utilized to reduce the amount of labeled data required for training, which is particularly beneficial when labeled data is scarce or expensive to obtain.

5.2. Findings on Machine Learning Algorithm Selection in Human Digital Twins

In Table 5, a list of the relevant articles found for questions 2 and 3 can be found. A total of 34 articles were found that discuss HDTs or aspects of personalized systems. Seven articles were found that addressed performance issues for HDTs. The articles listed in the general subcategory have been discussed in the previous sections. For the articles in the other three subcategories, an overview of the algorithms discussed, the datasets used, and the most relevant findings can be found in Tables A1–A4 in Appendix A.

Table 5. Articles found on personalization and Human Digital Twins.

Category		Publications
Machine Learning and Human Digital Twins	general	[4–6,12–14,16,22,24,35,36,38]
	single case example	[11,42,49,56,64–67]
	personalization	[21,23,28,37,50,53,68–75]
	performance	[15,28,48,53,68,76–78]

As described in Section 3, HDTs generally consist of several components or layers. In Figure 2, we can see that ML solutions can be implemented in most of these layers. The choices for the data acquisition or storage layer will, in general, be made by examining the options provided by algorithms suited for the data category. For the user-interaction layer, the most appropriate approach is found in the solutions supplied by RL-type algorithms. The implementation of the computation component is determined by the kind of modeling used in HDTs, the modeling of the human body or organ, or the modeling of behavior [12], but also by the objective of the computation component, as defined in Table 4.

The focus in this Section, and the next one, lies on the use of solution-type algorithms in HDTs. Section 5.4 will focus on the data-type algorithms in HDT implementations. In Table 6, several examples of HDT or healthcare applications and their corresponding problem types are given.

Table 6. Problem types from the solution categories and examples in HDTs.

Type	Example
Classification	Condition score in fitness management [11], emotion recognition in healthcare [42,65], user behavior modeling [53], disease classification [48,77], diagnose heart disease and detect heart problems [56,73,76], coaching applications [21]
Anomaly Detection	Ischemic heart diseases and stroke detection [64], anomalous behavior detection for elderly care [70]
Regression	Diagnosis and progression of cancer [66], metabolism models [68]
Reinforcement Learning	Personalized medicine prescription [28], personalized health interventions [72], coaching applications [21]
Clustering	Osteoarthritis endotype discovery [74], identifying biomarkers of lung health [75], unsupervised ECG analysis [49], personalized coaching applications [78]

In [11], a team of HDTs is used to predict the physical condition of participants during training for a team of athletes. Their work illustrates the challenges encountered when dealing with human data, specifically missing data, and much attention is therefore paid to data imputation. In [56,64,66], examples of human body or organ twinning are given. As is often the case, training data is usually based on patient databases and is therefore less personalized. As can be seen from the examples in this table, all types of problems within the solution category find applications in HDTs or healthcare settings.

Clustering examples, in general, involve USL models applied to data obtained from large groups of patients, used to identify patient groups. These models have traditionally not been used in personalized HDT applications; however, in [50], Trezza et al. provide an

overview of recent developments in precision medicine, an essential aspect of personalized healthcare, using USL.

In Table 7, the algorithms that are evaluated in the examples given in Table 6 are listed. The source of the model indicates whether a single model was generated based on the entire population (Group), multiple models were generated based on data from individual participants (Individual), or multiple models were generated based on clustering or a specific selection criterion (Other).

Table 7. Application objective, algorithms, and personalization level.

Category		Algorithms	Model Source
Anomaly Detection	[64]	CNN.	Group
	[70]	Wide range of classification techniques.	N/A
Classification	[21]	ADA, DTC, KNN,LR, NN, RF, SGD, SVC.	Group and Individual
	[77]	CNN, RF.	Group
	[48]	NB, KNN, RF.	Group
	[53]	ZeroR, NB, DTC, RF, SVM, KNN, ADA, LR, RIPPER, RIDOR, ANN.	Individual
	[42]	RF, CNN, K-Means.	Group
	[11]	KNN, SVM.	Individual
	[56]	RF, LR, SVM, C5.0.	Group
	[76]	CNN, LSTM, MLP, SVC, LR.	Group
	[73]	LR.	4 Age Groups
	[65]	Bagged trees, KNN, SVM (Linear and Cubic).	Group
Clustering	[49]	Wide range of unsupervised methods.	N/A
	[75]	K-Means.	Group
Clustering and Classification	[74]	K-Means, followed by RF	Group and 3 Clusters
	[78]	Manual clustering and RF.	3 Clusters
Regression	[68]	SARIMAX, LSTM, GRU, Transformer.	Group
	[66]	LR, DTR, RFR, GBA.	Group
Reinforcement Learning	[72]	Q-learning, LSPI, combined with K-medoids clustering.	Group, Clusters, and Individual

The selection of algorithms in the listed articles is, in most cases, based on general guidelines or expected potential to perform well. The conclusions in most articles are limited to presenting the algorithm that has the best performance for the given use case, without providing insight into the underlying reasons. Only [53,76] compare a wider selection of algorithms. Whereas the conclusions in [76] are restricted to the presented use-case, ref. [53] does reflect on the results. This study demonstrates that tree-based models yield more accurate predictions for context-aware smartphone usage models. In comparison, neural network-based models do not achieve the same prediction accuracy. One of the reasons suggested in this article is the limited number of samples in individual phone usage data.

Apart from our work in [21], no studies were found that compare algorithm selection or performance using individual and group models. However, in their discussion on the use of clustering techniques to generate labels used for classification in anomaly detection,

Nezambadi et al. [49] state that *Patient-specific ECG classifiers—trained classifiers that are fine-tuned over the ECG of the given patient—have shown superior performance over classifiers trained on a common ECG pool*, confirm our findings.

In conclusion, we can say that, based on the type of problem, a first selection of applicable machine learning models and algorithms can be made. Together with the learning model paradigm that must be chosen based on the availability of labels and information in the data, as described in Table 6, a complete classification of ML algorithms can be achieved. This classification then provides possible choices for algorithm selection, which can be refined by considering whether the models are generated using group data or individual data.

5.3. Findings on Human Digital Twin Performance Evaluation

Because humans are generally an active part of the system as patients or participants, one of the concerns that arises is that trust and motivation to participate play a crucial role in the data collection process, as well as in how output is received and responded to [15].

Motivational and trust issues can arise when the user experience is not optimal. The quality of this experience is determined by the choices made in the HDT's user interface layer. A good example of a survey with a more specific focus on ML implementations in this context is the work by den Hengst et al. [28]. Among other things, they concluded that further comparison of RL algorithms is needed and that the development of standardized personalization datasets and simulation environments is a promising direction for future research.

Another concern with personalized healthcare that must be taken into account is that, unlike industrial applications of DT, the physical twin itself is a cognitive being that interacts dynamically with its environment. The object of the DT is itself a learning subject, changing its behavioral patterns over time, which can lead to model or data drift. This means that adaptability and flexibility are essential considerations when designing and implementing HDTs.

For both of these concerns, evaluating the algorithm's performance can be part of the mitigation actions. This will enable the improvement of the accuracy of suggested interventions, calculated predictions, or the provision of more accurate feedback. Several studies were found that discussed the methods and results of such evaluations. A good example can be found in [68], where the implementation of a personalized metabolic avatar is described. In this work, the performance and computational time of four different models—SARIMAX, GRU, LSTM, and Transformer—were evaluated. Based on the experiment's results, a candidate for implementation in the production environment is suggested. An interesting point that they made in their discussion of the results is that some of the models exhibited significant variability among users. A related issue that could not be addressed in their study, due to the limited number of participants, was whether it would be more effective to have different models for different groups within the population or to train one model for all data.

In Table 8, an overview is given of the evaluation metric applied in the discussed articles. As can be seen from these results, only three articles discuss computational or execution time. In [48,68], the authors look at the performance of the selected algorithms measured at execution time, but no significant conclusions are drawn. In [78], the difference in model generation time between a clustered and a personalized approach is evaluated, and a substantial reduction in computational time is obtained.

Table 8. Application objective and evaluation metric.

Category		Evaluation Metric
Anomaly Detection	[64]	Accuracy.
	[70]	Accuracy, precision, recall and F-scores.
Classification	[21]	Accuracy and F1-score of classification using generalized or personal models.
	[77]	Confusion matrix.
	[48]	Accuracy, precision, recall, F-measures, and execution time.
	[53]	Precision, recall, F-measures, kappa, CCI, ICI, ROC, MAE, and RMSE.
	[42]	Accuracy.
	[11]	Classification loss.
	[56]	Accuracy, sensitivity, specificity, precision, AUC.
	[76]	Most metrics applicable to classification.
	[73]	F1-score, AUC and accuracy.
	[65]	Accuracy, precision, recall and F1-score.
Clustering	[49]	Accuracy.
	[75]	N/A
Clustering and Classification	[74]	Cluster stability by expanding the number of clusters from 3 to 5.
	[78]	Computational time, accuracy and F1-score.
Regression	[68]	RMSE and computational time.
	[66]	MSE, RMSE, MAE and R2.
Reinforcement Learning	[72]	Cumulative reward.

Several studies apply clustering or grouping techniques, and in general, the results from this approach are positive. For example, in [72], the authors state *A cluster-based RL can learn a significantly better policy within 100 days compared to learning per user and learning across all users, provided that a suitable clustering is found.*

These observations, along with the findings in the previous section, are consistent with the results of other studies identified in our survey. Little attention is paid to performance evaluation at the individual level, and the results obtained are generally based on calculations made on the entire dataset. In some cases, training or evaluation was based on data from a single participant; however, in all cases, it remains unclear whether different results could be obtained by examining individual subjects. No truly personalized approach to the evaluation of the ML algorithm can be found, but when clustering approaches are applied, promising results are obtained.

5.4. Findings on Human Digital Twin Dataset Considerations

A significant concern in HDT implementations is the quality and quantity of data. In Table 9, the 23 articles that discussed dataset considerations can be found. A recurring theme in the studies we examined is the issue of missing data. For example, in [11], a study on improving data quality by data imputation can be found. In the work of Fuller et al. [2], when discussing the challenges in obtaining useful data, it is stated that *"It needs to be quality data that are noise-free with a constant, uninterrupted data stream."* Since in HDTs some of the data, such as daily surveys, depends on the consistency of the participants, such quality requirements are challenging to meet [79]. Some of these concerns are also highlighted in the survey by Ding et al. [9] on the datasets used in human behavior analysis.

Table 9. Articles found on dataset considerations.

Category		Publications
Dataset considerations	General	[2,9,11,79–84]
	Feature evaluation	[70,85–91]
	Meta-learning	[61,62,78,92–94]

In Table 10, an overview of the data characteristics of the applications described in the articles in the previous sections is presented. We use these articles because most of the articles in Table 9 do not describe HDT applications. In Appendix A, more details can be found concerning the type of data and the content of the features.

Table 10. Application objective and dataset characteristics.

Category		Individuals	Features	Frequency	Data Type
Anomaly Detection	[64]	200	Signal	Stream	Physiological
	[70]	N/A	N/A	Stream	Physiological+Behavior
Classification	[21]	48	4	Periodical	Behavior
	[77]	257	52	Single	Physiological
	[48]	400	25	Single	Physiological
	[53]	10	N/A	Periodical	Behavior
	[42]	10	Image	Single	Physiological
	[11]	11	22	Periodical	Behavior+Context
	[56]	48	3	Stream	Physiological
	[76]	48	Signal	Stream	Physiological
	[73]	N/A	8	Single	Physiological+Behavior
	[65]	17	94	Stream	Behavior+Context
Clustering	[49]	2 to 500	Signal	Stream	Physiological
	[75]	656	51	Single	Physiological
Clustering and Classification	[74]	297	16	Single	Physiological
	[78]	43	4	Periodical	Behavior
Regression	[68]	10	5	Periodical	Physiological+Behavior
	[66]	116	11	Single	Physiological+Behavior
Reinforcement Learning	[72]	100	N/A	Periodical	Behavior

By examining the frequencies and data types of the applications listed, we can observe that, in general, DTs that focus on disease detection and diagnosis utilize physiological data, such as MRI scans (single observations) or ECG data (streams). DTs that perform coaching tasks rely on behavioral data such as activity data or step data, which can be complemented by context or physiological data. By examining the number of patients or participants, we observe that in most cases, the number of individuals in each dataset is limited. However, the number of records per individual can be quite extensive. In general, the number of individuals and the resulting data quantity are not considered problematic; however, some articles suggest that the results could be different or improved if more individual datasets were available [53,68]. Data quality is discussed in most articles, but only [11] provides a detailed discussion on the measures taken to improve data quality.

Another problem with human data, identified by Deep et al. [70], is caused by confounding variables and activities. In some cases, activities that are very different can yield similar data readings from sensors, such as smartwatches or IoT devices. This is called confounding activities. Confounding variables arise when an unmeasured variable influences several others, creating apparent causal connections. These phenomena make

feature selection an essential part of improving the efficiency and accuracy of the learning task. In the article by Deep et al., one of the suggestions made for further research on algorithm improvement is to investigate the implementation of deep learning models such as DNN, CNN, Autoencoder, RBM, RNN, and independent component analysis algorithms for automatic feature extraction and classification.

Feature selection is a well-studied field in the context of data science, particularly when analyzing large datasets with a substantial number of features. In our survey, we found that feature selection is considered crucial to making the learning process more efficient and precise. It can be challenging to determine the most suitable algorithm from the many available and *"often relies on the expertise of a human or a random trial-and-error approach"* [86].

In the context of HDTs, the challenge often lies in the fact that the number of features is relatively large compared to the number of samples. In [87], an approach is proposed to achieve meaningful information for feature selection in this situation. When the sample size is larger, one of the techniques used is meta-learning, where the characteristics of the dataset are used as input for selection processes [94].

Most often, meta-learning is employed in preprocessing and feature selection, as well as in algorithm selection, particularly when addressing classification problems. In [93], a method is proposed to apply meta-learning to recommend the most suitable classification algorithm for a given situation. In this method, both structural and statistical information are used to identify the nearest neighbors within a labeled set of datasets. Based on these nearest neighbors, the appropriate classification algorithm is recommended. The results of this study demonstrate that meta-learning can be effective in selecting classification algorithms. However, in our research, we have not found an application of meta-learning in the context of HDTs. In our recent work [78], we discuss the possibility of clustering-based meta-learning for the VFC platform as a first step to explore the opportunities that meta-learning offers in the context of HDTs.

6. Conclusions, Perspectives, and Discussion

We discussed that one of the significant advantages of HDT is its ability to provide personalized support in healthcare. Creating efficient, personalized healthcare through the use of HDTs presents several challenges. Data quality and quantity can be an issue, as can noise and confounding variables. Furthermore, considering that the physical twin is a cognitive entity that adapts to its environment and the Digital Twin's output demands constant monitoring of the accuracy of its virtual counterpart.

As can be seen from a comparison between the findings on ML and HDT and the findings on datasets, there is a sharp contrast between the approach taken to feature selection and, in a lesser way, to algorithm selection in the field of data science, compared to the way selection processes are conducted for HDT applications. Although studies have been conducted for nearly two decades on the subject of meta-learning using dataset characteristics, there seems to be a total absence of its application in HDTs. Since data quality and algorithm performance are crucial factors in the success of HDTs, and in many cases, regular model updates are necessary to compensate for drift, we identify opportunities for improvement in this area.

Furthermore, we have established that there is a lack of knowledge regarding the selection and evaluation of personalized models. The most common approach is to use generalized models, and there are no truly personalized assessment methods. This is essential, considering individual adaptation, to keep the virtual object optimally adapted for the physical object in the HDT configuration.

Further Research

Based on the discussion in the previous section, we suggest the following research question for further research.

Question. *How can the optimization of human data feature engineering and personalized model selection be achieved in Human Digital Twins, and can techniques such as meta-learning be of use in this context?*

The focus here should lie on investigating the challenges that arise from the fact that each human, and thus each physical twin in the Digital Twin system, is a unique, evolving individual, requiring continuous personalized adaptation of the ML components of the HDT application. By answering this question, more insight should be gained for ways to improve HDT's accuracy. Higher precision is essential for better diagnosis results, used, for example, for precision medicine. Furthermore, improved behavior prediction, utilized in a wide range of coaching and supervision applications, will enhance the trust and motivation of patients or participants using these applications.

Author Contributions: Conceptualization, H.H.R., P.C.-C. and M.T.; methodology, H.H.R. and P.C.-C.; software, H.H.R.; validation, P.C.-C., T.B.D., H.K.E.O. and M.T.; formal analysis, H.H.R., T.B.D. and P.C.-C.; investigation, H.H.R.; resources, T.B.D.; data curation, H.H.R. and T.B.D.; writing—original draft preparation, H.H.R.; writing—review and editing, P.C.-C., T.B.D., H.K.E.O. and M.T.; visualization, H.H.R.; supervision, P.C.-C., H.K.E.O. and M.T.; project administration, H.K.E.O. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Not applicable.

Abbreviations

The following abbreviations are used in this manuscript:

DT	Digital Twin
HDT	Human Digital Twin
AI	Artificial Intelligence
ML	Machine Learning
IoT	Internet of Things
VFC	Virtual Fitness Coach
SL	Supervised Learning
USL	Unsupervised Learning
SSL	Semi-Supervised Learning
RL	Reinforcement Learning
AUC	Area under the curve
CCI	correctly classified instances
ICI	incorrectly classified instances
RMSE	Root Mean Square Error
MSE	Mean Squared Error
MAE	Mean Absolute Error
ROC	Receiver Operating Characteristics
R2	Coefficient of Determination

ADA	AdaBoost Classifier
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DTC	Decision Tree Classifier
DTR	Decision Tree Regression
ESN	Echo State Network
FCN	Fully Convolutional Neural Networks
GBA	Gradient Boost Algorithm
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Units
KNN	k-Neighbors Classifier
LR	Logistic Regression Classifier
LSPI	Least-Squares Policy Iteration
LSTM	Long Short-Term Memory networks
MLP	Multi Layer Perceptron
NB	Naive Bayes
NLR	Non-linear regression
NN	Neural Networking Classifier
RBM	Restricted Boltzmann machine
RF	Random Forest Classifier
RFR	Random Forest Regression
RIDOR	Ripple Down Rule Learner
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RNN	Recurrent Neural Network
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors
SGD	Stochastic Gradient Descent Classifier
SMOTE	Synthetic Minority Over-sampling Technology
SVC	Support Vector Classifier
WNN	Wavelet Neural Networks

Appendix A. List of Features

Table A1. A chronological overview of the articles found on algorithm selection and performance evaluation for HDTs.

Publication Date	Dataset Type Dataset Size	Algorithms Evaluation Method/Metric	Results
[21] 2018-02	Step data collected over 12 weeks, resulting in 349,920 measurements.	ADA, DTC, KNN, LR, NN, RF, SGD, SVC.	On average, the personalized models outperform the general models, with the RF algorithm achieving the highest average accuracy of 93%. But there is a significant spread in optimal model when looking at an individual level.
	48 participants.	Accuracy and F1-score of classification using generalized or personal models.	
[71] 2018-07	N/A	N/A	This survey focuses on the different wearable devices available for self-health tracking.
[77] 2018-07	Per patient 275 feature maps based on MRI scans were created, of which 52 were selected.	CNN, RF.	In this study, an RF-based segmentation method is compared to more commonly used CNN models and shown to provide good results.
	257 patients.	Confusion matrix.	
[72] 2018-10	Simulated activity data with 3 different user profiles.	Q-learning, LSPI, combined with K-medoids clustering.	This study compares batch learning to online learning but additionally shows that a clustering approach can achieve good results compared to individual and non-personalized learning approaches.
	Simulated data for 100 users.	Cumulative reward.	
[48] 2019-03	25 features describing blood values.	NB, KNN, RF.	RF classifier performs best on the given dataset, but there was only a single point of evaluation, and no investigation of time development.
	400 instances.	Accuracy, precision, recall, F-measures, and execution time.	
[53] 2019-07	Ten phone log datasets, with 55,105 phone call activities and metadata.	ZeroR, NB, DTC, RF, SVM, KNN, ADA, LR, RIPPER, RIDOR, ANN.	The results of this study show that tree-based models yield higher prediction results for context-aware smartphone usage models. In comparison, neural network-based models do not achieve the same prediction accuracy. One reason for this could be the limited number of samples in the individual phone usage data.
	10 individuals.	Precision, recall, F-measures, kappa, CCI, ICI, ROC, MAE and RMSE.	
[64] 2019-08	ECG data stream.	CNN.	In this article, a proof of concept is examined. Although the authors discuss the options of other models that the CNN model used, no evaluation of other models is performed.
	Data from 200 patients.	Accuracy.	

Table A2. A chronological overview of the articles found on algorithm selection and performance evaluation for HDTs.

Publication Date	Dataset Type Dataset Size	Algorithms Evaluation Method/Metric	Results
[42] 2019-12	Physical information such as gender, blood pressure, and cholesterol was used with the RF classifier. Facial image data was used with K-Means and 2 CNN classifiers.	RF, CNN, K-Means.	No evaluation or justification of the selected algorithms is given.
	10 participants.	Accuracy.	
[70] 2020-01	Various datasets concerning anomalous behavior detection for elderly care.	Wide range of classification techniques.	This survey presents a list of pros and cons for investigated methods but no explicit comparison.
	Various.	Accuracy, precision, recall and F-scores.	
[11] 2020-02	22 features, containing athletes' activity, mood, and energy intake data.	KNN, SVM.	SVM classifiers are more robust but achieve worse performance than KNN classifiers.
	11 participants, 10 measurements consisting of data over 3 days per participant, resulting in 110 data vectors.	Classification loss.	
[28] 2020-04	N/A	RL algorithms.	This survey provides an overview of work that employs RL for personalization.
		N/A	
[69] 2020-04	MRI scan.	N/A	The article proposes an architecture for a digital twin of the behavior of lung cancer in patients but provides no technical details.
	N/A		
[37] 2020-10	N/A	SVM, CNN, KNN, Trees, RNN, LR, GMM, NN, LSTM, ESN, WNN, non-linear regression, FCN, NB.	This survey gives an overview of the different models used in current studies, and also the types of wearable devices used.
		N/A	
[67] 2020-10	N/A	N/A	This study describes the architecture for a DT aimed at providing precision medicine for MS patients.

Table A3. A chronological overview of the articles found on algorithm selection and performance evaluation for HDTs.

Publication Date	Dataset Type Dataset Size	Algorithms Evaluation Method/Metric	Results
[56] 2021-12	Electroencephalography, one channel vertical electro-oculogram and one channel chin electro-myogram.	RF, LR, SVM, C5.0.	The study validated the proof of concept of a Digital Twin using an EEG headset. The SVM classifier was found to achieve the highest accuracy, but no in-depth evaluation of the different algorithms was provided.
	Data from 48 stroke survivors.	Accuracy, sensitivity, specificity, precision, AUC.	
[76] 2021-12	MIT-BIH Arrhythmia Database.	CNN, LSTM, MLP, SVC, LR.	The metrics of the different classifiers are compared with each other, showing that NN-based classifiers outperformed the others for some metrics, but the LSTM classifier had the best results for macro and weighted average for precision, recall, and F1-score.
	48 half-hour excerpts of two-channel ambulatory ECG recordings.	Most metrics applicable to classification.	
[49] 2022-02	Electrocardiogram data.	Wide range of unsupervised methods.	Several clustering techniques are compared, and different applications of the clustering results are discussed. The different algorithms are compared, but only general use cases are suggested. No specific selection is proposed.
	Ranging between 2 and 500.	Accuracy.	
[73] 2022-03	Age and 7 risk factors consisting of a combination of physiological data and behavior data.	LR.	The article focuses on the implementation of combining mechanical models with ML models and does not provide detailed information on the ML models used.
	N/A	F1-Score, AUC and accuracy.	
[74] 2022-03	16 biomarkers collected from serum and urine. Extra information, including assessment of radiographs of knees and hands, MRIs, and CT scans of the knees, and outcomes of physical examinations and questionnaires.	K-Means, followed by RF.	The combination of clustering followed by RF classification makes it possible to determine which variables determine the cluster membership.
	297 patients.	Cluster stability by expanding the number of clusters from 3 to 5.	
[15] 2022-05	N/A	N/A	In this survey, the distinction between DT and HDT is identified, and several additional design requirements are introduced.

Table A4. A chronological overview of the articles found on algorithm selection and performance evaluation for HDTs.

Publication Date	Dataset Type Dataset Size	Algorithms Evaluation Method/Metric	Results
[23] 2022-07	N/A	N/A	This study provides key design features and an architectural framework to implement HDTs. It also highlights some technical challenges.
[75] 2023-01	51 plasma EV-miRNAs.	K-Means.	In this study, clustering is used to identify lung health. The focus lies on the usability of the given dataset to achieve this goal, but it does not expand on the ML techniques used.
	656 patients.	N/A	
[68] 2023-02	Weight, activity and diet.	SARIMAX, LSTM, GRU, Transformer.	Based on computational time and the RMSE results, this study concludes that GRU or LSTM is best to be used in a production environment.
	10 participants, 100 days.	RMSE and computational time.	
[66] 2023-03	Blood samples, demographic, anthropometric, and clinical data.	LR, DTC, RFR, GBA.	Although the results for the individual models are presented by the authors, rather than selecting one optimal algorithm, the approach in this article is to combine the results of the four different models.
	116 participants, 52 healthy women and 64 with breast cancer.	MSE, RMSE, MAE, and R2.	
[65] 2023-06	Facial imaging and body movement data.	Bagged trees, KNN, SVM (linear and cubic).	The authors perform an in-depth analysis of the data structure and qualities of the models and conclude *Due to the nature of the data, the non-linear algorithms produced consistent findings. All the classification methods consistently performed worse than the Bagged Trees and k-NN.*
	17 participants, performing 6 emotional states.	Accuracy, precision, recall, and F1-score.	
[78] 2024-09	Step data.	Manual clustering and RF.	This study demonstrates that a clustering approach to personalization is a viable concept that can reduce computational time without significant loss of accuracy.
	43	Computational time, accuracy, and F1-score.	
[50] 2024-09	N/A	Several clustering, dimensionality reduction, and anomaly detection techniques.	This review demonstrates the applicability of UL techniques to improve precision medicine applications.
		Accuracy.	

References

1. Rosen, R.; von Wichert, G.; Lo, G.; Bettenhausen, K.D. About The Importance of Autonomy and Digital Twins for the Future of Manufacturing. *IFAC-PapersOnLine* **2015**, *48*, 567–572. [CrossRef]

2. Fuller, A.; Fan, Z.; Day, C.; Barlow, C. Digital Twin: Enabling Technologies, Challenges and Open Research. *IEEE Access* **2020**, *8*, 108952–108971. [CrossRef]

3. Barricelli, B.R.; Casiraghi, E.; Fogli, D. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access* **2019**, *7*, 167653–167671. [CrossRef]

4. Wang, B.; Zhou, H.; Yang, G.; Li, X.; Yang, H. Human Digital Twin (HDT) Driven Human-Cyber-Physical Systems: Key Technologies and Applications. *Chin. J. Mech. Eng. (Engl. Ed.)* **2022**, *35*, 11. [CrossRef]

5. Mihai, S.; Yaqoob, M.; Hung, D.V.; Davis, W.; Towakel, P.; Raza, M.; Karamanoglu, M.; Barn, B.; Shetve, D.; Prasad, R.V.; et al. Digital Twins: A Survey on Enabling Technologies, Challenges, Trends and Future Prospects. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 2255–2291. [CrossRef]

6. Guo, J.; Lv, Z. Application of Digital Twins in multiple fields. *Multimed. Tools Appl.* **2022**, *81*, 26941–26967. [CrossRef]

7. Adeniyi, A.O.; Arowoogun, J.O.; Okolo, C.A.; Chidi, R.; Babawarun, O. Ethical considerations in healthcare IT: A review of data privacy and patient consent issues. *World J. Adv. Res. Rev.* **2024**, *21*, 1660–1668. [CrossRef]

8. Rajput, D.; Wang, W.J.; Chen, C.C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinform.* **2023**, *24*, 48. [CrossRef]

9. Ding, X.; Gan, Q.; Bahrami, S. A systematic survey of data mining and big data in human behavior analysis: Current datasets and models. *Trans. Emerg. Telecommun. Technol.* **2022**, *33*, e4574. [CrossRef]

10. Tucker, A.; Wang, Z.; Rotalinti, Y.; Myles, P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digit. Med.* **2020**, *3*, 147. [CrossRef]

11. Barricelli, B.R.; Casiraghi, E.; Gliozzo, J.; Petrini, A.; Valtolina, S. Human Digital Twin for Fitness Management. *IEEE Access* **2020**, *8*, 26637–26664. [CrossRef]

12. Lin, Y.; Chen, L.; Ali, A.; Nugent, C.; Ian, C.; Li, R.; Gao, D.; Wang, H.; Wang, Y.; Ning, H. Human Digital Twin: A Survey. *arXiv* **2022**, arXiv:2212.05937. [CrossRef]

13. Shengli, W. Is Human Digital Twin possible? *Comput. Methods Programs Biomed. Update* **2021**, *1*, 100014. [CrossRef]

14. Alazab, M.; Khan, L.U.; Koppu, S.; Ramu, S.P.; M, I.; Boobalan, P.; Baker, T.; Maddikunta, P.K.R.; Gadekallu, T.R.; Aljuhani, A. Digital Twins for Healthcare 4.0—Recent Advances, Architecture, and Open Challenges. *IEEE Consum. Electron. Mag.* **2022**, *12*, 29–37. [CrossRef]

15. Lauer-Schmaltz, M.W.; Cash, P.; Hansen, J.P.; Maier, A. Designing Human Digital Twins for Behaviour-Changing Therapy and Rehabilitation: A Systematic Review. *Proc. Des. Soc.* **2022**, *2*, 1303–1312. [CrossRef]

16. Ahmadi-Assalemi, G.; Al-Khateeb, H.; Maple, C.; Epiphaniou, G.; Alhaboby, Z.A.; Alkaabi, S.; Alhaboby, D. Digital Twins for Precision Healthcare. In *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*; Jahankhani, H.; Kendzierskyj, S.; Chelvachandran, N.; Ibarra, J., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 133–158. [CrossRef]

17. Blok, J.; Dol, A.; Dijkhuis, T. Toward a Generic Personalized Virtual Coach for Self-management: A Proposal for an Architecture. In Proceedings of the 9th International Conference on eHealth, Telemedicine, and Social Medicine 2017, Hanze University of Applied Sciences, Nice, France, 19–23 March 2017.

18. Dijkhuis, T.; Blok, J.; Velthuijsen, H. Virtual Coach: Predict Physical Activity Using A Machine Learning Approach. In Proceedings of the eTELEMED 2018: The Tenth International Conference on eHealth, Telemedicine, and Social Medicine, Hanze University of Applied Sciences, Rome, Italy, 25–29 March 2018.

19. Schoeppe, S.; Alley, S.; Van Lippevelde, W.; Bray, N.A.; Williams, S.L.; Duncan, M.J.; Vandelanotte, C. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: A systematic review. *Int. J. Behav. Nutr. Phys. Act.* **2016**, *13*, 1–26. [CrossRef]

20. Hardeman, W.; Houghton, J.; Lane, K.; Jones, A.; Naughton, F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. *Int. J. Behav. Nutr. Phys. Act.* **2019**, *16*, 31. [CrossRef]

21. Dijkhuis, T.B.; Blaauw, F.J.; van Ittersum, M.W.; Velthuijsen, H.; Aiello, M. Personalized physical activity coaching: A machine learning approach. *Sensors* **2018**, *18*, 623. [CrossRef]

22. Chen, J.; Yi, C.; Okegbile, S.D.; Cai, J.; Shen, X. Networking Architecture and Key Supporting Technologies for Human Digital Twin in Personalized Healthcare: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2023**, *26*, 706–746. [CrossRef]

23. Okegbile, S.D.; Cai, J.; Yi, C.; Niyato, D. Human Digital Twin for Personalized Healthcare: Vision, Architecture and Future Directions. *IEEE Netw.* **2022**, *37*, 262–269. [CrossRef]

24. Liu, Y.; Zhang, L.; Yang, Y.; Zhou, L.; Ren, L.; Wang, F.; Liu, R.; Pang, Z.; Deen, M.J. A Novel Cloud-Based Framework for the Elderly Healthcare Services Using Digital Twin. *IEEE Access* **2019**, *7*, 49088–49101. [CrossRef]

25. Alzubi, J.; Nayyar, A.; Kumar, A. Machine Learning from Theory to Algorithms: An Overview. *J. Phys. Conf. Ser.* **2018**, *1142*, 012012. [CrossRef]

26. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef]

27. Mahesh, B. Machine learning algorithms—A review. *Int. J. Sci. Res. (IJSR)* **2020**, *9*, 381–386. [CrossRef]

28. den Hengst, F.; Grua, E.M.; el Hassouni, A.; Hoogendoorn, M. Reinforcement learning for personalization: A systematic literature review. *Data Sci.* **2020**, *3*, 107–147. [CrossRef]

29. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.F.; Hua, L. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *J. Med. Syst.* **2012**, *36*, 2431–2448. [CrossRef]

30. Semeraro, C.; Lezoche, M.; Panetto, H.; Dassisti, M. Digital twin paradigm: A systematic literature review. *Comput. Ind.* **2021**, *130*, 103469. [CrossRef]

31. Miller, M.E.; Spatz, E. A unified view of a human digital twin. *Hum.-Intell. Syst. Integr.* **2022**, *4*, 23–33. [CrossRef]

32. El Saddik, A. Digital Twins: The Convergence of Multimedia Technologies. *IEEE Multimed.* **2018**, *25*, 87–92. [CrossRef]

33. Kamali, M.E.; Angelini, L.; Caon, M.; Carrino, F.; Röcke, C.; Guye, S.; Rizzo, G.; Mastropietro, A.; Sykora, M.; Elayan, S.; et al. Virtual Coaches for Older Adults' Wellbeing: A Systematic Review. *IEEE Access* **2020**, *8*, 101884–101902. [CrossRef]

34. Bruynseels, K.; de Sio, F.S.; van den Hoven, J. Digital Twins in health care: Ethical implications of an emerging engineering paradigm. *Front. Genet.* **2018**, *9*, 31. [CrossRef] [PubMed]

35. Chatterjee, A.; Prinz, A.; Gerdes, M.; Martinez, S. Digital Interventions on Healthy Lifestyle Management: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e26931. [CrossRef] [PubMed]

36. De Maeyer, C.; Markopoulos, P. Are Digital Twins Becoming Our Personal (Predictive) Advisors?: 'Our Digital Mirror of Who We Were, Who We Are and Who We Will Become'. In Proceedings of the Lecture Notes in Computer Science (Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Copenhagen, Denmark, 19–24 July 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 250–268. [CrossRef]

37. Gámez Díaz, R.; Yu, Q.; Ding, Y.; Laamarti, F.; El Saddik, A. Digital Twin Coaching for Physical Activities: A Survey. *Sensors* **2020**, *20*, 5936. [CrossRef]

38. Minerva, R.; Lee, G.M.; Crespi, N. Digital Twin in the IoT Context: A Survey on Technical Features, Scenarios, and Architectural Models. *Proc. IEEE* **2020**, *108*, 1785–1824. [CrossRef]

39. Samuel, A.L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229. [CrossRef]

40. Mitchell, T.M.T.M. *Machine Learning*; McGraw-Hill Science: New York, NY, USA, 1997; pp. 1–432.

41. Huijzer, R.; Blaauw, F.; den Hartigh, R.J.R. SIRUS.jl: Interpretable Machine Learning via Rule Extraction. *J. Open Source Softw.* **2023**, *8*, 5786. [CrossRef]

42. Abeydeera, S.S.; Bandaranayake, M.; Karunarathna, H.U.; Pallewatta, S.; Dharmasiri, P.; Gunathilake, B.; Saparamadu, S.; Senanayake, B.; Jayawardena, C. Smart Mirror with Virtual Twin. In Proceedings of the 2019 International Conference on Advancements in Computing, ICAC 2019, Malabe, Sri Lanka, 5–7 December 2019; pp. 238–243. [CrossRef]

43. Bouchlaghem, Y.; Akhiat, Y.; Amjad, S. Feature Selection: A Review and Comparative Study. In Proceedings of the E3S Web of Conferences, Istanbul, Turkey, 12–14 May 2022; EDP Sciences: Paris, France, 2022; Volume 351. [CrossRef]

44. Lötsch, J.; Ultsch, A. Enhancing Explainable Machine Learning by Reconsidering Initially Unselected Items in Feature Selection for Classification. *BioMedInformatics* **2022**, *2*, 701–714. [CrossRef]

45. Han, J.; Kamber, M.; Pei, J. Cluster Analysis: Basic Concepts and Methods. *Data Min.* **2012**, 443–495. [CrossRef]

46. Min, Q.; Lu, Y.; Liu, Z.; Su, C.; Wang, B. Machine Learning based Digital Twin Framework for Production Optimization in Petrochemical Industry. *Int. J. Inf. Manag.* **2019**, *49*, 502–519. [CrossRef]

47. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]

48. Devika, R.; Avilala, S.V.; Subramaniyaswamy, V. Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest. In Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 27–29 March 2019; pp. 679–684. [CrossRef]

49. Nezamabadi, K.; Sardaripour, N.; Haghi, B.; Forouzanfar, M. Unsupervised ECG Analysis: A Review. *IEEE Rev. Biomed. Eng.* **2022**, *16*, 208–224. [CrossRef] [PubMed]

50. Trezza, A.; Visibelli, A.; Roncaglia, B.; Spiga, O.; Santucci, A. Unsupervised Learning in Precision Medicine: Unlocking Personalized Healthcare through AI. *Appl. Sci.* **2024**, *14*, 9305. [CrossRef]

51. Chatterjee, A.; Pahari, N.; Prinz, A.; Riegler, M. Machine learning and ontology in eCoaching for personalized activity level monitoring and recommendation generation. *Sci. Rep.* **2022**, *12*, 19825. [CrossRef] [PubMed]

52. Wu, N.; Green, B.; Ben, X.; O'Banion, S. Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case. *arXiv* **2020**. [CrossRef]

53. Sarker, I.H.; Kayes, A.S.; Watters, P. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *J. Big Data* **2019**, *6*, 57. [CrossRef]

54. Jianshan Sun Zhiqiang Tian, Y.F.J.G.; Liu, C. Digital twins in human understanding: A deep learning-based method to recognize personality traits. *Int. J. Comput. Integr. Manuf.* **2021**, *34*, 860–873. [CrossRef]

55. Lee, M.C.; Lin, J.C.; Gan, E.G. ReRe: A Lightweight Real-Time Ready-to-Go Anomaly Detection Approach for Time Series. In Proceedings of the 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 13–17 July 2020; pp. 322–327. [CrossRef]

56. Hussain, I.; Hossain, M.A.; Park, S.J. A Healthcare Digital Twin for Diagnosis of Stroke. In Proceedings of the 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 4–5 December 2021, pp. 18–21. [CrossRef]

57. Villamizar, H.; Kalinowski, M.; Lopes, H.; Mendez, D. Identifying concerns when specifying machine learning-enabled systems: A perspective-based approach. *J. Syst. Softw.* **2024**, *213*, 112053. [CrossRef]

58. Hancer, E.; Xue, B.; Zhang, M. A survey on feature selection approaches for clustering. *Artif. Intell. Rev.* **2020**, *53*, 4519–4545. [CrossRef]

59. Yakovyna, V.; Shakhovska, N.; Szpakowska, A. A novel hybrid supervised and unsupervised hierarchical ensemble for COVID-19 cases and mortality prediction. *Sci. Rep.* **2024**, *14*, 9782. [CrossRef]

60. Singh, A.; Thakur, N.; Sharma, A. A review of supervised machine learning algorithms. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 1310–1315.

61. Monteiro, J.P.; Ramos, D.; Carneiro, D.; Duarte, F.; Fernandes, J.M.; Novais, P. Meta-learning and the new challenges of machine learning. *Int. J. Intell. Syst.* **2021**, *36*, 6240–6272. [CrossRef]

62. Brazdil, P.; Gama, J.; Henery, B. Characterizing the applicability of classification algorithms using meta-level learning. *Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **1994**, *784*, 83–102. [CrossRef]

63. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2023**, *622*, 178–210. [CrossRef]

64. Martinez-Velazquez, R.; Gamez, R.; El Saddik, A. Cardio Twin: A Digital Twin of the human heart running on the edge. In Proceedings of the 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Istanbul, Turkey, 26–28 June 2019; pp. 1–6. [CrossRef]

65. Amara, K.; Kerdjidj, O.; Ramzan, N. Emotion Recognition for Affective Human Digital Twin by Means of Virtual Reality Enabling Technologies. *IEEE Access* **2023**, *11*, 74216–74227. [CrossRef]

66. Moztarzadeh, O.; Jamshidi, M.B.; Sargolzaei, S.; Jamshidi, A.; Baghalipour, N.; Malekzadeh Moghani, M.; Hauer, L. Metaverse and Healthcare: Machine Learning-Enabled Digital Twins of Cancer. *Bioengineering* **2023**, *10*, 455. [CrossRef]

67. Petrova-Antonova, D.; Spasov, I.; Krasteva, I.; Manova, I.; Ilieva, S. A digital twin platform for diagnostics and rehabilitation of multiple sclerosis. In Proceedings of the Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, 1–4 July 2020; pp. 503–518. [CrossRef]

68. Abeltino, A.; Bianchetti, G.; Serantoni, C.; Riente, A.; De Spirito, M.; Maulucci, G. Putting the Personalized Metabolic Avatar into Production: A Comparison between Deep-Learning and Statistical Models for Weight Prediction. *Nutrients* **2023**, *15*, 1199. [CrossRef]

69. Angulo, C.; Gonzalez-Abril, L.; Raya, C.; Ortega, J.A. A Proposal to Evolving Towards Digital Twins in Healthcare. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Granada, Spain, 6–8 May 2020; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12108 LNBI, pp. 418–426. [CrossRef]

70. Deep, S.; Zheng, X.; Karmakar, C.; Yu, D.; Hamey, L.G.C.; Jin, J. A Survey on Anomalous Behavior Detection for Elderly Care Using Dense-Sensing Networks. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 352–370. [CrossRef]

71. Dias, D.; Cunha, J.P.S. Wearable Health Devices—Vital Sign Monitoring, Systems and Technologies. *Sensors* **2018**, *18*, 2414. [CrossRef]

72. el Hassouni, A.; Hoogendoorn, M.; van Otterlo, M.; Barbaro, E. Personalization of Health Interventions using Cluster-Based Reinforcement Learning. In Proceedings of the PRIMA 2018: Principles and Practice of Multi-Agent Systems: 21st International Conference, Tokyo, Japan, 29 October–2 November 2018; pp. 467–475. [CrossRef]

73. Herrgårdh, T.; Hunter, E.; Tunedal, K.; Örman, H.; Amann, J.; Navarro, F.A.; Martinez-Costa, C.; Kelleher, J.D.; Cedersund, G. *Digital Twins and Hybrid Modelling for Simulation of Physiological Variables and Stroke Risk*; Cold Spring Harbor Laboratory: Laurel Hollow, NY, USA, 2022.

74. Angelini, F.; Widera, P.; Mobasheri, A.; Blair, J.; Struglics, A.; Uebelhoer, M.; Henrotin, Y.; Marijnissen, A.C.; Kloppenburg, M.; Blanco, F.J.; et al. Osteoarthritis endotype discovery via clustering of biochemical marker data. *Ann. Rheum. Dis.* **2022**, *81*, 666–675. [CrossRef]

75. Eckhardt, C.M.; Gambazza, S.; Bloomquist, T.R.; De Hoff, P.; Vuppala, A.; Vokonas, P.S.; Litonjua, A.A.; Sparrow, D.; Parvez, F.; Laurent, L.C.; et al. Extracellular Vesicle-Encapsulated microRNAs as Novel Biomarkers of Lung Health. *Am. J. Respir. Crit. Care Med.* **2023**, *207*, 50–59. [CrossRef]

76. Elayan, H.; Aloqaily, M.; Guizani, M. Digital Twin for Intelligent Context-Aware IoT Healthcare Systems. *IEEE Internet Things J.* **2021**, *8*, 16749–16757. [CrossRef]

77. Bonte, S.; Goethals, I.; Van Holen, R. Machine learning based brain tumour segmentation on limited data using local texture and abnormality. *Comput. Biol. Med.* **2018**, *98*, 39–47. [CrossRef] [PubMed]

78. Van Buren, A.; Kwan, A.; Rietdijk, H.H.; Dijkhuis, T.B.; Conde-Cespedes, P.; Oldenhuis, H.; Trocan, M. A Clustering Approach for Personalized Coaching Applications. In Proceedings of the Advances in Computational Collective Intelligence, Leipzig, Germany, 9–11 September 2024; Nguyen, N.-T., Franczyk, B., Ludwig, A., Treur, J., Vossen, G., Kozierkiewicz, A., Eds.; pp. 351–363. [CrossRef]

79. Konsolakis, K.; Banos, O.; Cabrita, M.; Hermens, H. COVID-BEHAVE dataset: Measuring human behaviour during the COVID-19 pandemic. *Sci. Data* **2022**, *9*, 754. [CrossRef] [PubMed]

80. Tatti, N. Distances between Data Sets Based on Summary Statistics. *J. Mach. Learn. Res.* **2007**, *8*, 131–154.

81. Banaee, H.; Ahmed, M.U.; Loutfi, A. Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. *Sensors* **2013**, *13*, 17472–17500. [CrossRef]

82. Park, Y.; Ho, J.C. Tackling Overfitting in Boosting for Noisy Healthcare Data. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 2995–3006. [CrossRef]

83. Liu, F.; Demosthenes, P. Real-world data: A brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* **2022**, *22*, 287. [CrossRef]

84. Singh, A.; Halgamuge, M.N.; Lakshmiganthan, R. Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*. [CrossRef]

85. Oh, S. A new dataset evaluation method based on category overlap. *Comput. Biol. Med.* **2011**, *41*, 115–122. [CrossRef]

86. Parmezan, A.R.S.; Lee, H.D.; Spolaôr, N.; Wu, F.C. Automatic recommendation of feature selection algorithms based on dataset characteristics. *Expert Syst. Appl.* **2021**, *185*, 115589. [CrossRef]

87. Rietdijk, H.H.; Strijbos, D.O.; Conde-Cespedes, P.; Dijkhuis, T.B.; Oldenhuis, H.K.E.; Trocan, M. Feature Selection with Small Data Sets: Identifying Feature Importance for Predictive Classification of Return-to-Work Date after Knee Arthroplasty. *Appl. Sci.* **2024**, *14*, 9389. [CrossRef]

88. Oreski, D.; Oreski, S.; Klicek, B. Effects of dataset characteristics on the performance of feature selection techniques. *Appl. Soft Comput.* **2017**, *52*, 109–119. [CrossRef]

89. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform.* **2022**, *2*, 927312. [CrossRef] [PubMed]

90. Remeseiro, B.; Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **2019**, *112*, 103375. [CrossRef]

91. Wang, L.; Han, M.; Li, X.; Zhang, N.; Cheng, H. Review of Classification Methods on Unbalanced Data Sets. *IEEE Access* **2021**, *9*, 64606–64628. [CrossRef]

92. Al Masud, A.; Hossain, S.; Rifa, M.; Akter, F.; Zaman, A.; Farid, D.M. Meta-Learning in Supervised Machine Learning. In Proceedings of the 2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Phnom Penh, Cambodia, 2–4 December 2022; pp. 222–227. [CrossRef]

93. Song, Q.; Wang, G.; Wang, C. Automatic recommendation of classification algorithms based on data set characteristics. *Pattern Recognit.* **2012**, *45*, 2672–2689. [CrossRef]

94. Pio, P.B.; Rivolli, A.; Carvalho, A.C.P.L.F.d.; Garcia, L.P.F. A review on preprocessing algorithm selection with meta-learning. *Knowl. Inf. Syst.* **2023**, *66*, 1–28. [CrossRef]